О дискретных аппроксимациях непрерывных вероятностных распределений

Фрей Александр Ильич

Московский физико-технический институт
(Государственный университет)
Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы» (ВЦ РАН)

Научный руководитель: к.ф.-м.н. Воронцов Константин Вячеславович

18 июня 2008

Основные определения

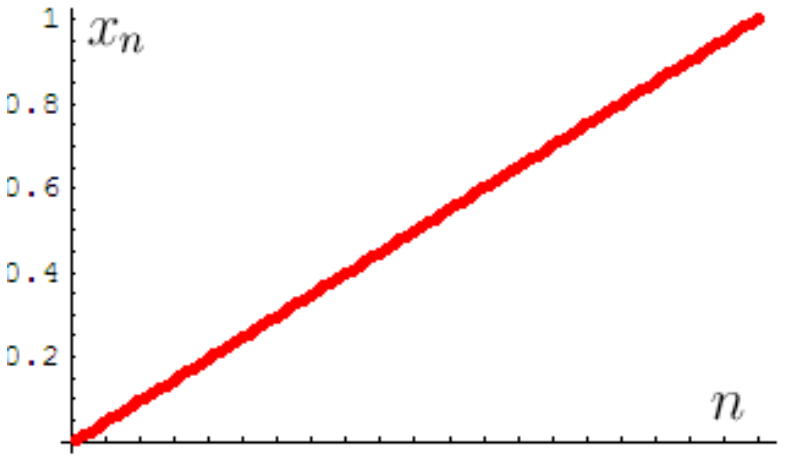
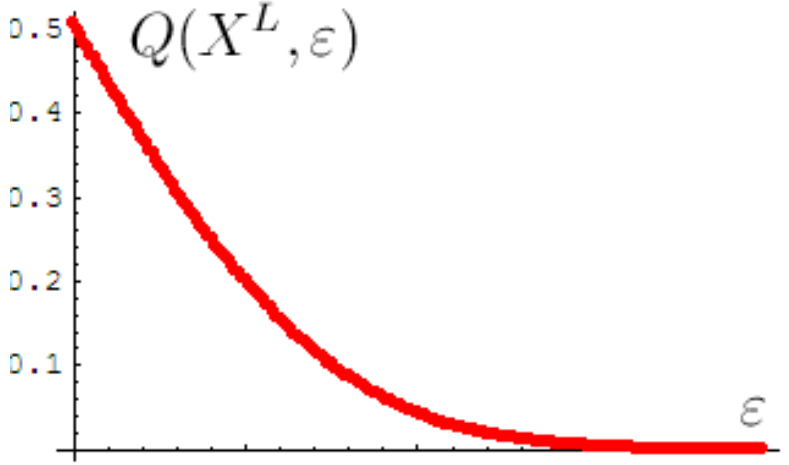
- **Выборка** — набор из L чисел:
  $X^L = \{x_1, x_2, \ldots, x_L\} \in \mathbb{R}^L$
- $X^L = X_n^k \cup X_n^\ell$, где $n \in 1, \ldots, N; N = C_L^k$ — всевозможные разбиения на наблюдаемую и скрытую подвыборки.
- **Отклонение средних** в двух подвыборках —

$$D(X^k, X^\ell) = \frac{1}{k} \sum_{x \in X^k} x - \frac{1}{\ell} \sum_{x \in X^\ell} x$$

- **Вероятность больших уклонений** —

$$Q(X^L, \varepsilon) = \frac{1}{N} \sum_{n=1}^{N} \Big[D(X_n^k, X_n^\ell) \geq \varepsilon \Big],$$

Цели исследования

Вариационный ряд X^L	Зависимость $Q(X^L, \varepsilon)$ от ε
	

Цели исследования:

- получить оценку $Q(X^L, \varepsilon) \leq Q(Y^L, \varepsilon)$, где Y^L — выборка из дискретного множества (например, $\{0, 1\}$).
- исследовать непрерывные распределения комбинаторными методами

Классические верхние оценки

Пусть $X_1, \ldots, X_n$ — независимые случайные величины. Обозначим $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{D}(X_i)$.

- Неравенство Чебышева:

$$\mathrm{P}(|S_n - \mathrm{E}(S_n)| \geq \varepsilon) \leq \frac{\sigma_n^2}{n\varepsilon^2}$$

- Неравенство Чебышева-Кантелли

$$\mathrm{P}(S_n - \mathrm{E}(S_n) \geq \varepsilon) \leq \frac{\sigma_n^2}{\sigma_n^2 + n\varepsilon^2}$$

- Неравенство Хёфдинга

$$\mathrm{P}(S_n - \mathrm{E}(S_n) \geq \varepsilon) \leq e^{-2n\varepsilon^2}$$

- Неравенство Бернштейна

$$\mathrm{P}(S_n - \mathrm{E}(S_n) \geq \varepsilon) \leq \exp\left(-\frac{2n\varepsilon^2}{4\sigma_n^2 + \varepsilon}\right)$$

Результаты в слабой вероятностной аксиоматики

- Для выборки $Y_m^L \in \{0,1\}^L$, состоящей из нулей и единиц, выполнена точная оценка

$$Q(Y_m^L, \varepsilon) = \sum_{t=s_0}^{s_1} h(\tbinom{\ell\, t}{L\, m}),$$

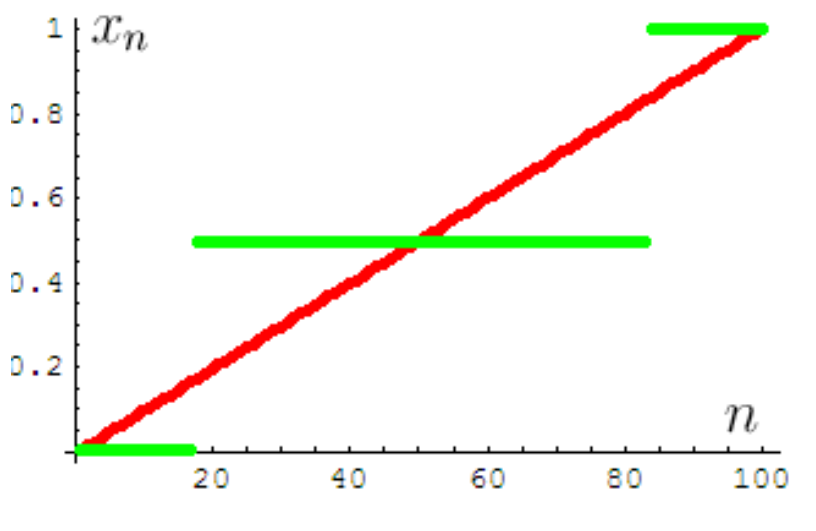
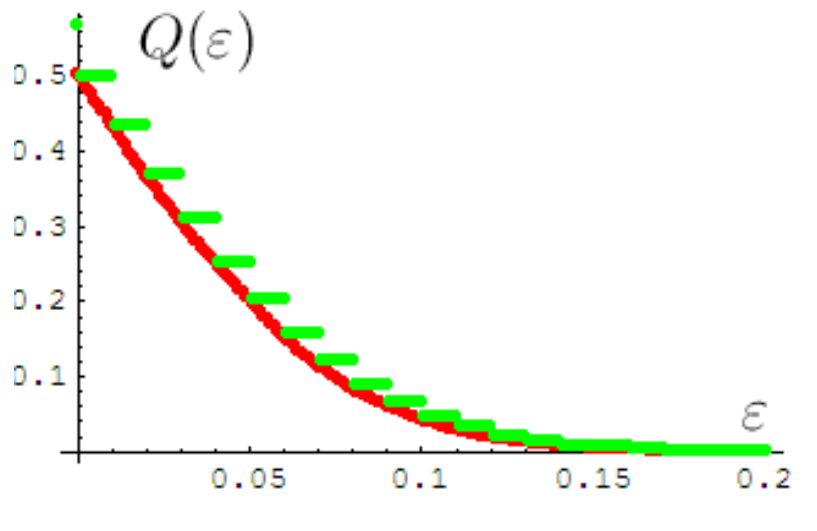
где m - число единиц в выборке, $s_0 = \max(0, m-k)$, $s_1 = \lfloor (m - \varepsilon k)\,\frac{\ell}{L} \rfloor$, $h(\tbinom{\ell\, t}{L\, m}) = \dfrac{C_{L-m}^{\ell-t} C_m^t}{C_L^\ell}$.

Теорема (А. Бадзян)

При $k = \ell$ $\forall X^L \in [0,1]^L$, $\forall \varepsilon > 0$ $\exists m \in \{0, \ldots, L\}$, *такое что*

$$Q(X^L, \varepsilon) \le Q(Y_m^L, \varepsilon).$$

Трехступенчатая выборка Z^L

Вариационный ряд X^L	Зависимость $Q(X^L, \varepsilon)$ от ε
	

- Выборка Z^L из $\{0, q, 1\}$, число значений $n_0 + n_q + n_1 = L$.

Теорема (Вероятность больших уклонений для выборки Z^L)

$$Q(Z^L, \varepsilon) = \frac{1}{N} \sum_{\ell_q=0}^{n_q} C_{n_q}^{\ell_q} \sum_{\ell_1=0}^{s_1} C_{L-n_q-n_1}^{\ell-\ell_q-\ell_1} C_{n_1}^{\ell_1}, \qquad (1)$$

где $s_1 = \left\lfloor \frac{\ell(qn_q+n_1)-k\ell\varepsilon}{L} - \ell_q q \right\rfloor$.

Метод моментов. Теорема адекватности

- Первые моменты: $\mu_1 = \sum x_i$, $\mu_2 = \sum x_i^2$, $\mu_3 = \sum x_i^3$.
- Метод моментов: для выбора параметров q, n_q, n_1 приравняем первые три момента выборки X^L и Z^L

Теорема (Адекватности)

Решение уравнений метода моментов имеет вид

$$\begin{cases} q = \dfrac{\mu_2 - \mu_3}{\mu_1 - \mu_2} \\ n_q^* = \dfrac{(\mu_1 - \mu_2)^3}{(\mu_2 - \mu_3)(\mu_1 - 2\mu_2 + \mu_3)} \\ n_1^* = \dfrac{\mu_1\mu_3 - \mu_2^2}{\mu_1 - 2\mu_2 + \mu_3} \end{cases}$$

и удовлетворяет естественным условиям

$$0 \le q \le 1,\ n_0^* \ge 0,\ n_q^* \ge 0,\ n_1^* \ge 0.$$

Тонкости округления

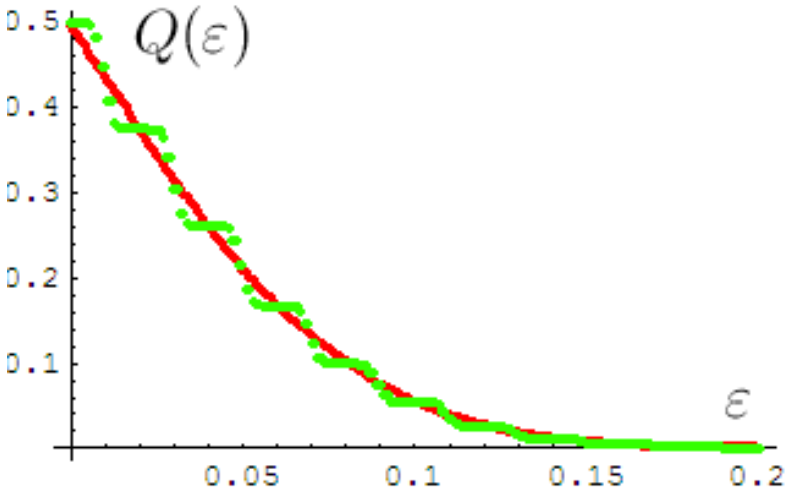
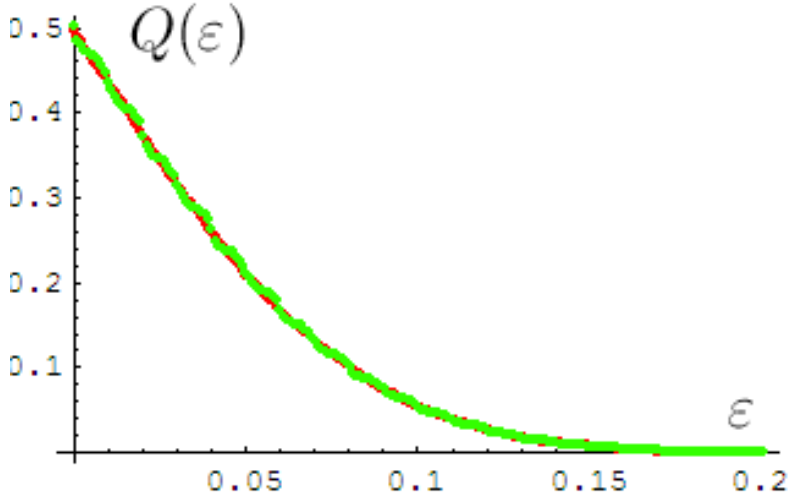
- Введем параметр r - размер окрестности округления.
- **Окрестность округления** $B_r(n_0^*, n_q^*, n_1^*)$ — все тройки неотрицательных целых чисел (n_0, n_q, n_1), получаемые из (n_0^*, n_q^*, n_1^*) изменением каждой координаты не более чем на r, при условии $n_0 + n_q + n_1 = L$.

Теорема (О количестве точек в окрестности округления)

Пусть r является целым числом, и все параметры задачи (n_0^*, n_q^*, n_1^*) далеко (т.е. не менее чем на r) отстоят от граничных значений 0 и L и не являются целыми. Тогда размер окрестности простым образом выражается через ее радиус:

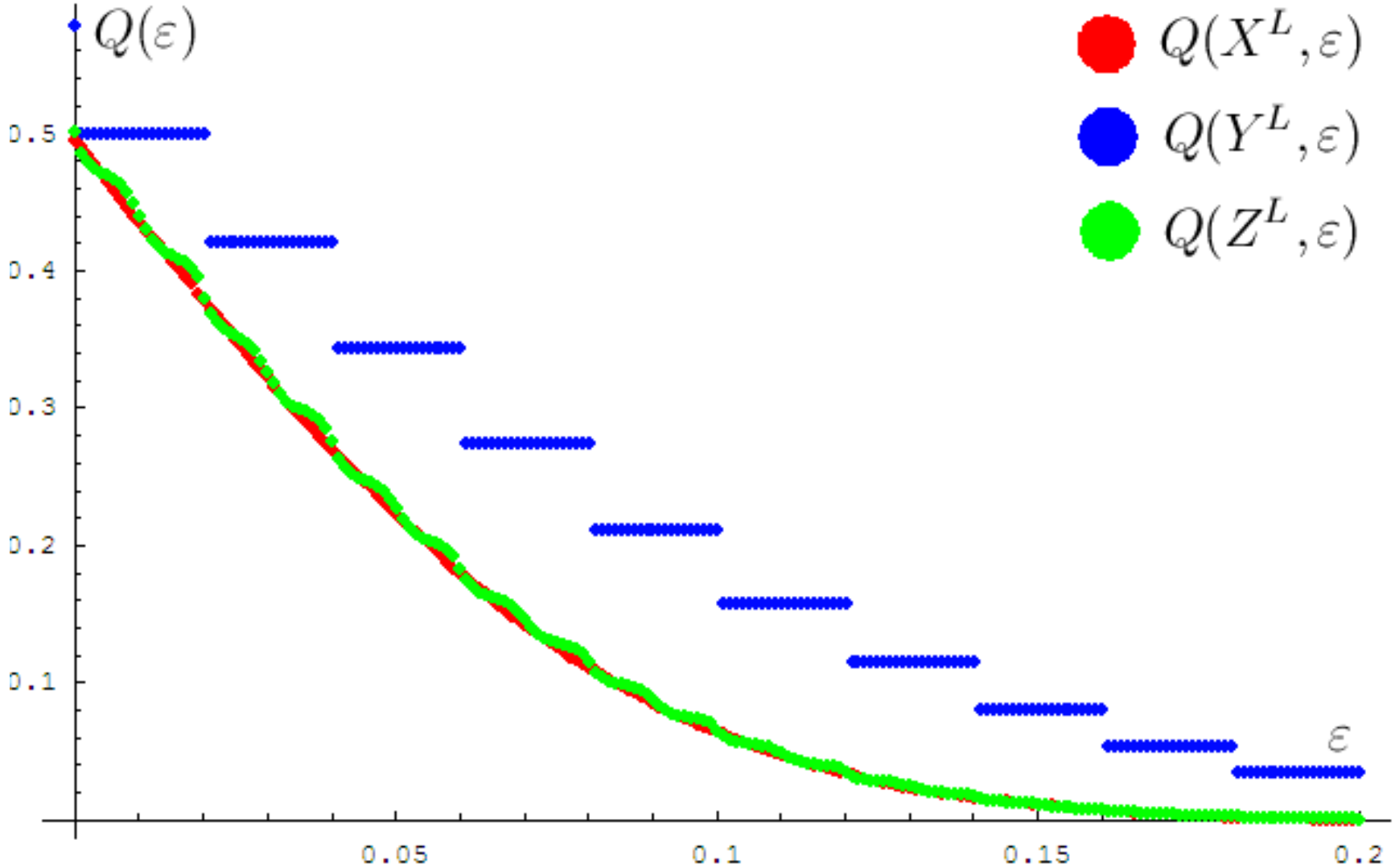
$$|B_r| = 3r^2.$$

Контрпримеры

До усреднений	После усреднения
	

- Существуют контрпримеры, показывающие что даже при округлении по окрестности радиуса $r = 2$ оценка не является верхней.
- Предлагается усреднить выражение $Q(Z^L, \varepsilon)$ по всем $(n_0, n_q, n_1) \in B_r(n_0^*, n_q^*, n_1^*)$.

Численные результаты - сравнение оценок



Основные результаты и направления исследований

- Получены **комбинаторные** оценки для непрерывных распределений
- Получено явное выражение для вероятности больших уклонений трехступенчатой выборки Z^L
- Предложен метод моментов для вычисления параметров выборки и доказана его корректность
- Изучены тонкости округления, присущие данной задаче
- Вычислено количество точек, приближающих решение метода моментов

Направления дальнейших исследований:

- Улучшение оценки $Q(X^L, \varepsilon) \leq Q(Y_m^L, \varepsilon)$ выбором числа m
- Получение строгой верхней оценки $Q(X^L, \varepsilon) \leq Q(Z^L, \varepsilon)$
- Получение асимптотических результатов при больших L и оценка скорости сходимости